An Improved Multiband Spectral Subtraction Technique (with Perceptual Post Filtering) for Speech Enhancement

Conference Paper · December 2003

DOI: 10.13140/2.1.5175.4564

CITATIONS
2

READS
573

4 authors, including:

Chetak Kandaswamy
Flemish Institute for Technological Research
15 PUBLICATIONS **298** CITATIONS

SEE PROFILE

M. Girish Chandra
Tata Consultancy Services Limited
144 PUBLICATIONS **837** CITATIONS

SEE PROFILE

An Improved Multiband Spectral Subtraction Technique (with Perceptual Post Filtering) for Speech Enhancement

Karthik Venkat*, Chetak Kandaswamy*, M. Girish Chandra+, Gurumurthy*

*University Visvesvaraya College of Engineering, Bangalore
+National Aerospace Laboratories, Bangalore

Abstract

An Improved Multiband Spectral Subtraction (IMBSS) for speech enhancement is considered in this paper. Unlike the conventional MBSS, the band-wise processing is used in pre-processing, spectral flooring and noise updating also, apart from the subtraction process. Additionally, the band-subtraction factors are made adaptive using a novel quantity called band energy content. Also, the noise spectrum is updated at certain frequency bands within a frame along with the conventional methods of noise updating during silent frames. All these are incorporated to have an algorithm, which can provide a greater control of noise removal properties in different frequency bands. A perceptual post filter to suppress only the noise audible to human ear is incorporated into the proposed method to further improve the enhancing capability of the technique. Apart from the details relevant to the said issues, the paper presents some typical results to demonstrate performance improvement in terms of objective measures and subjective listening tests, offered by the proposed technique compared to the conventional method.

Introduction

The performance of voice communication systems degrades rapidly in adverse acoustic environments. When a speaker and listener communicate in a quiet environment, information exchange is easy and accurate. But, this is not true in many speech communication settings, as they are invariably associated with background noise or interference. Examples include machine noise in a factory environment, car noise in automobiles, cafeteria noise, etc. Further, though speech can be perceived in a moderately noisy environment [1], in many sophisticated applications, the presence of the background noise is very undesirable as it results in overall loss of intelligibility and gives an uncomfortable feeling of listener fatigue. Examples include mobile communications, aids for the hearing impaired [2] and cockpit voice recorders (CVR) to name a few. In short, degradation of speech severely affects the ability of a person to understand what the speaker is saying and curtails the performance of the communication system. To combat the above problems Speech Enhancement (SE) techniques are being increasingly employed. SE will improve the quality and/or intelligibility of corrupted speech. Improving speech quality generally refers to reduction of unwanted noise resulting in improved signal to noise (SNR) ratio of processed speech. Additionally, the SE techniques should distort the speech as little as possible during the process of noise reduction, thus providing a good perceptual quality speech after the processing.

Apart from the applications mentioned above, SE techniques are being recognized as invaluable front-end tools for other allied speech processing applications like speech recognition, speech coding (compression), competing speaker suppression (multi-speaker babble).

The Spectral Subtraction (SS) proposed by Boll [3], is one of the earliest and by far the most popular SE technique. The basic principle of SS is to subtract noise from that of the noisy speech. An estimate of the noise signal is obtained during silence or non-speech activity in the speech signal. Though SS is simple and easy to implement it has some shortcomings. A prominent one being the presence of musical noise [3], which at times is more annoying than the original disturbance. To reduce the musical noise and provide better overall performance, several SE techniques have been developed [4] from the extension of SS, resulting in a class of subtractive-type algorithms. The Multi-band Spectral Subtraction (MBSS) algorithm [5] is one such technique that was developed mainly to address the problems posed by real-world noise (colored noise) in speech. It exploits a frequency-dependent subtraction by using band-specific over-subtraction factor. The results obtained by MBSS provide notable improvements over the SS and its variants. However, the MBSS also has some shortcomings (elaborated later) like the values for the band specific factors have to be obtained empirically.

In this paper, an effort has been made to carry out certain modifications to the MBSS method to achieve superior speech quality and intelligibility while causing minimal distortions to underlining speech. The proposed method, which will be referred to as Improved MBSS (IMBSS) introduces a new parameter called as the Band Energy Content (BEC) to obtain values of the band specific factors. Also, as an offshoot of this, an experiment has been made to update the noise spectrum in a band-wise manner, which is not a common practice.

The use of perceptual properties for SE is a recent significant step in the research community ([11], [12]). Thus, a perceptual post filter (PPF) based on human auditory masking has been incorporated in the IMBSS method. Performance evaluation showed that the proposed method provides significant improvements in terms of objective measure like SNRs and Itakura-Saito (IS) distance measure [14], [15]. Further, informal listening tests showed that more speech components were retained by this method (i.e. minimal distortions to underlying

speech) with the elimination of musical noise. A distinguishing feature of IMBSS is that while MBSS was aimed at mainly for colored noise conditions, the proposed method performs well for both white as well as colored noise.

To present IMBSS and relevant results, the rest of the paper is organized as follows. In Section 2, the modeling of noise and speech commonly used in subtractive type algorithms, followed by an outline of the SS and MBSS techniques is given. Section 3 captures the details of IMBSS method. The perceptual post-filtering portion of IMBSS is considered in a cursory fashion in Section 4. The results and discussion of the work are presented in Section 5.

2. SS and MBSS

The discussion in this paper is confined to single microphone situation, where noise and speech are present in the same channel (single channel case). As usual, the noise signal $d(\mathrm{n})$ $d(n)$ is assumed to be an additive stationary random process, which is uncorrelated with the speech signal $s(n)$, resulting in the corrupted (noisy) speech $y(n)$:

$$y(n) = s(n) + d(n) \quad (1)$$

As in majority of speech processing techniques, short-term frames (using a window) are considered in enhancement also. That is, the noisy speech signal is split into short-term frames using the window $w(n)$ as:

$$y_{pL}(n) = w(pL - n)\big(s(n) + d(n)\big) \quad (2)$$

where, L is the frame (window) length and p is an integer, signifying the frame number (p th frame). In this paper, Hamming window is used, consistent with the majority of the speech-enhancement works. When Eqn.1 is considered in the frequency domain, the following equation results:

$$Y(pL,\omega) = S(pL,\omega) + D(pL,\omega) \quad (3)$$

where, $Y(pL,\omega)$, $S(pL,\omega)$ and $D(pL,\omega)$ are the Short-term Fourier Transform (STFT) of the noisy speech $y(n)$, clean speech $s(n)$ and background noise $d(n)$, respectively. The STFT magnitude squared of $y(n)$ can be written as [6]:

$$|Y(pL,\omega)|^2 \approx |S(pL,\omega)|^2 + |D(pL,\omega)|^2 \quad (4)$$

For implementation purposes, Discrete Fourier Transform (DFT) is used, and the DFT $Y(pL,k)$ can be written in the familiar amplitude and phase form as

$$Y(pL,k) = |Y(pL,k)|e^{i\phi(pL,k)} \quad (5)$$

2.1 Spectral Subtraction (SS) [3]

The spectral subtraction technique stems from a simple observation of Eqn.4. If an estimate of the noise power spectrum $|\hat{D}(k)|^2$ is available, then an estimate of the speech (clean speech) spectrum can be obtained as

$$|\hat{S}(pL,k)|^2 = |Y(pL,k)|^2 - |\hat{D}(pL,k)|^2 \quad (6)$$

Since the noise spectrum $|D(k)|^2$ cannot be directly obtained, the power spectrum $|\hat{D}(k)|^2$ is estimated during a period of silence [7], using time average.

From Eqn.6 it can be clearly seen that the subtraction process involves the subtraction of an averaged estimate of the noise from the instantaneous speech spectrum. Due to the differences in the values of the estimated noise spectrum and actual (instantaneous) noise spectrum (which is unavoidable), one may end up with some negative values for the enhanced spectrum. These values are set to zero by a process of half-wave rectification [3] and can be written as:

$$|\hat{S}(pL,k)|^2 = \begin{cases} |\hat{S}(pL,k)|^2 & \text{if } |\hat{S}(pL,k)|^2 > 0 \\ 0 & \text{else} \end{cases} \quad (7)$$

The modified spectrum of Eqn.7 is combined with the phase information from the noise-corrupted signal to reconstruct the time-domain signal by using the Inverse Discrete Fourier Transform (IDFT)

$$\hat{s}_{pL}(n) = IDFT\left(|\hat{S}(pL,k)|e^{j\phi(pL,k)}\right) \quad (8)$$

The estimate of the complete clean speech $\hat{s}(n)$ is obtained from adding the individual frames $\hat{s}_{pL}(n)$, thus taking any overlap also into consideration. This method is usually referred to as the overlap and add (OLA) method [6].

While the spectral subtraction method is easily implemented and effectively reduces the noise present in the corrupted signal, there exist some glaring shortcomings, like the presence of residual noise (musical noise), roughening of phase, etc [5].

The presence of musical noise is reasoned in the following way. Since there can be some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum, the subtraction of these quantities results in the presence of isolated residual noise levels of large variance. This residual spectral content manifests themselves in the reconstructed time signal as varying tonal sounds, resulting in musical noise. This musical noise can be even more disturbing and annoying to the listener than the distortions due to the original noise content. Over the years many additional techniques were suggested using the principle of SS to come up with methods to reduce musical noise. Some of the methods are application specific and/or condition specific (like colored noise case).

2.2 Variations of Spectral Subtraction [5][7]

Several variants of the spectral subtraction method originally proposed by Boll have been developed to address the problems of the basic technique. Two of them

are (1) Magnitude Averaging and (2) Spectral Subtraction using over subtraction and spectral floor. The first method reduces spectral error by averaging across neighboring (spectral) frames. This has the effect of lowering the noise variance while reinforcing the speech spectral content and thus preventing destructive subtraction. The second method is an important variation of spectral subtraction for reduction of residual (musical) noise. This technique could be expressed as [7]:

$$|\hat{S}(pL,k)|^2 = |Y(pL,k)|^2 - \alpha |\hat{D}(pL,k)|^2 \quad (8)$$

$$|\hat{S}(pL,k)|^2 = \begin{cases} |\hat{S}(pL,k)|^2 & \text{if } |\hat{S}(pL,k)|^2 > \beta |\hat{D}(pL,k)|^2 \\ \beta |\hat{D}(pL,k)|^2 & \text{else} \end{cases} \quad (9)$$

where, the over-subtraction factor α is a function of the signal-to-noise ratio (SNR). The over-subtraction factor subtracts an overestimate of the noise power spectrum from the speech power spectrum. This operation minimizes the presence of residual noise by decreasing the spectral excursions in the enhanced spectrum.

2.3 Multi-Band Spectral Subtraction (MBSS) [5]

Most implementations and variations of the SS technique advocate subtraction of noise spectrum estimate over the entire speech spectrum. But real world noise is mostly colored and does not affect speech signal uniformly over entire spectrum. To exploit the non-uniform effects of noise on speech, a frequency-dependent spectral subtraction approach is proposed in [5] called Multi-Band Spectral Subtraction (MBSS). In this method, the frequency band is split into $N\ (1 \le N \le 8)$ portions and spectral subtraction is performed independently on each band using band-specific over-subtraction factors. This introduces non-linearity for the subtraction procedure (non-linear subtraction procedures are quite recently examined by research community). The steps involved in the MBSS method are as follows:

1. Windowing and FFT:

The speech signal is divided into frames of 20ms using Hamming window with 10ms overlap. For each frame a N-point DFT is taken and corresponding representation in frequency domain is obtained.

2. Preprocessing:

To bring the instantaneous noise content close to mean noise spectrum (i.e. reduce noisy speech spectral variance) and hence reduce the perception of residual noise in enhanced speech, a smoothed version of the power spectra of the signal is used instead of the signal spectra directly. Preprocessing consists of local magnitude averaging.

$$\bar{Y}(pL,k) = \sum_{l=-M}^{M} W_l\, Y((p-l)L,k) \quad (10)$$

where, p is the frame index and W_l is weighting factor, $0 < W_l < 1$. The averaging is done over M preceding and succeeding frames (Eqn.10) of speech resulting in $\bar{Y}(pL,k)$, which is the smoothed and averaged version of the noisy speech spectrum. Then each frame is divided into N uniform bands.

3. Noise Spectrum Estimate:

Estimated noise spectrum $|\hat{D}(pL,k)|$ is calculated during silent zones. However, initially it is assumed that the first frame is silent and its magnitude spectrum gives $|\hat{D}(pL,k)|$, which is used for the subsequent subtraction and updated as and when new silent zones are detected (see step 6).

4. Spectral Subtraction:

The power spectrum estimate of the clean speech spectrum in the ith band of the p^{th} frame can be written as:

$$|\hat{S}_i(pL,k)|^2 = |\bar{Y}_i(pL,k)|^2 - \alpha_i\, \delta_i\, |\hat{D}_i(pL,k)|^2 \quad b_i \le k \le e_i \quad (11)$$

where, α_i is the over-subtraction factor of the ith band, δ_i is a band-subtraction factor that can be individually set for each frequency band to customize the noise removal properties and b_i, e_i denote the beginning and ending frequencies of ith frequency band.

The band specific over-subtraction factor α_i used in (11) is a function of segmental SNR_i of the ith frequency band. SNR_i is calculated as:

$$SNR_i(dB) = 10 \log_{10} \left(\frac{\sum_{k=b_i}^{e_i} |Y_i(pL,k)|^2}{\sum_{k=b_i}^{e_i} |\hat{D}_i(pL,k)|^2} \right) \quad (12)$$

The over-subtraction factor α_i as a function of segmental SNR is shown in Fig.1.

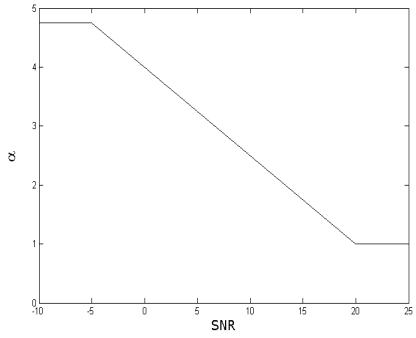


Fig 1: Variation of α with respect to SNR for $\alpha_0 = 4$

In the figure, α_0 is the desired value of α_i at 0 dB. For $\alpha_0 = 4$, using above graph we have the following condition.

$$\alpha_i = \begin{cases} 4.75 & SNR_i < -5 \\ 4 - (3/20)(SNR_i) & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \quad (13)$$

Another important factor on which the performance of subtraction process defined in Eqn.11 depends is the band-subtraction factors δ_i. The values for this are set individually for each frequency band. The values of δ_i could be varied in order to emphasize or de-emphasize to particular frequency band. Thus, the noisy removal properties have to be customized with values of δ_i. The values of δ_i used in [5] are:

$$\delta_i = \begin{cases} 1 & f_i \le 1kHz \\ 2.5 & 1kHz < f_i \le \frac{Fs}{2} - 2kHz \\ 1.5 & f_i > \frac{Fs}{2} - 2kHz \end{cases} \quad (14)$$

where, f_i denotes upper frequency of ith band and F_s sampling frequency (in Hz).

5. Spectral Flooring

Due to the error in computing the noise spectrum, one may have some negative values in the modified spectrum. The negative values obtained in the Eqn.13 are floored as:

$$|\hat{S}(pL,k)|^2 = \begin{cases} |\hat{S}(pL,k)|^2 & \text{if } |\hat{S}(pL,k)|^2 > 0 \\ \beta |\hat{D}(pL,k)|^2 & \text{otherwise} \end{cases} \quad (15)$$

the spectral floor β prevents the spectral components of the enhanced spectrum from falling below the lower value, $\beta |\hat{D}(pL,k)|^2$ and $\beta = 0.002$.

6. Noise Updating (Silence Detection):

This is accomplished using a Voice Activity Detector (VAD). Given a previous value of estimated noise spectrum, $|\hat{D}(pL,k)|$ (see step 3), the likelihood ratio of speech being present or absent in the current p^{th} input frames is detected by evaluating the following condition:

$$\mu = \frac{1}{N}\sum_{k=1}^{N-1}\left(\frac{|Y_i(pL,k)|^2}{|\hat{D}_{i-1}(pL,k)|^2} - \log_{10}\frac{|Y_i(pL,k)|^2}{|\hat{D}_{i-1}(pL,k)|^2} - 1 \right) \quad (16)$$

If $\mu > \eta$, where η is the preset threshold, then speech is present and if $\mu \le \eta$ then speech is absent, which implies that only noise is present (silent frame). If the non-speech activity is detected in a p^{th} frame then $|\hat{D}(pL,k)|$ for that frame is updated as

$$|\hat{D}_j(pL,k)|^2 = \lambda_d \cdot |\hat{D}_{j-1}(pL,k)|^2 + (1-\lambda_d) \cdot |Y_j(pL,k)|^2 \quad (17)$$

and the same is used for subtraction during the subsequent frames until a next silent zone is detected. In (17), λ_d is the decaying factor (and can be set to 0.9 [5]).

7. IDFT & Overlap and Add (OLA):

The N bands of a frame are recombined back to obtain one single frame $Y(pL,k)$. IDFT is computed for every frame using original noisy phase. The frames are combined to obtain the estimate of the original signal using OLA method.

The MBSS method, though reduces background noise levels substantially, some shortcomings are observable. These are elaborated in the next section.

3. IMBSS Algorithm

This Section begins with the motivating factors for the proposed method followed by details of the algorithm.

3.1 Motivation for the Proposed Algorithm

In the MBSS method, the values of band-subtraction factors δ_i are individually set for each frequency band for customizing noise removal properties in that band, as in Eqn.14. These values have to be empirically assigned by placing certain constraints on noise/speech parameters or by trial and error method. Furthermore, thus obtained values may not be accurate, which may result in over or under subtractions. Thus, the method lacks a proper methodology to calculate these factors which are crucial to the output performance. Additionally, since the values of δ_i remain constant throughout the subtraction process, it is difficult for the system to respond to the fast changing noise/speech spectral properties.

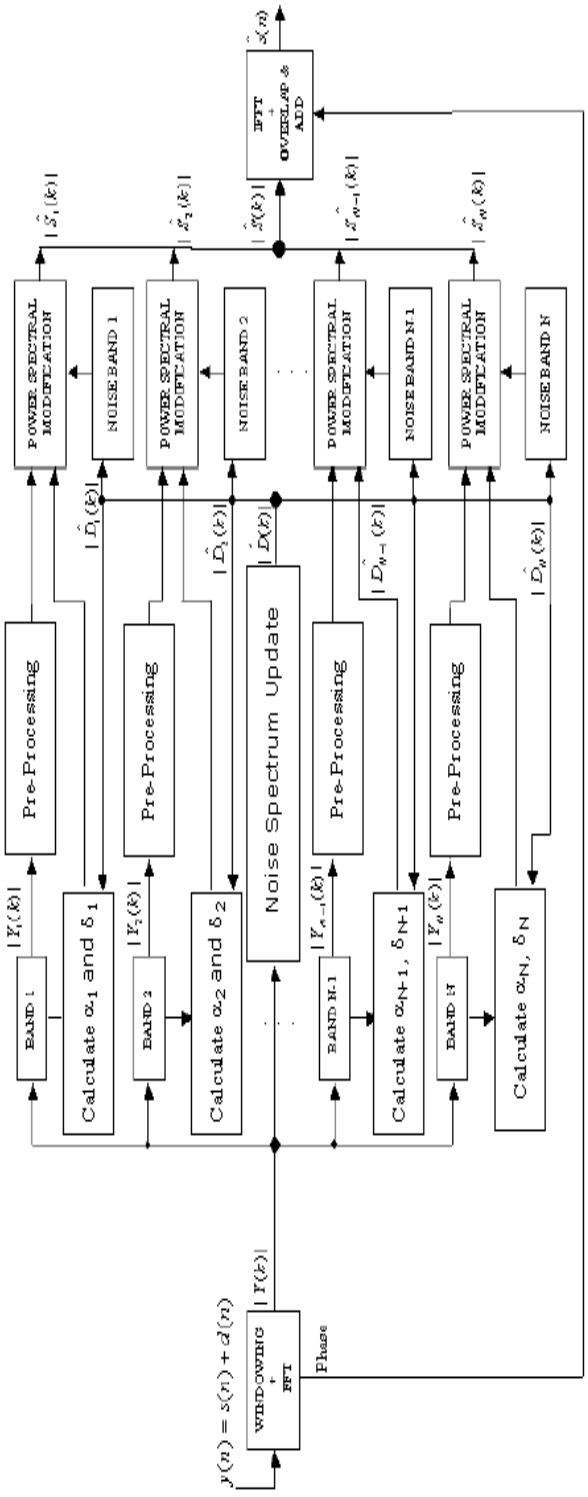
From the careful observations of the listening tests carried out for the MBSS method, it is noted that the overall speech intelligibility and naturalness of processed speech is reduced. This method, apart from removing background noise, introduces some distortions to the underlying speech, which is highly objectionable. Due to incomplete subtraction some residual noise is left which manifests as musical noise.

To address these drawbacks and enhance the performance of the MBSS method, we propose a new method called as Improved Multi-Band Spectral Subtraction (IMBSS) technique. As the output of any SE algorithm has to be ultimately judged by the human ear, use of perceptual properties based on human auditory system for SE has become an active area of research ([6], [11], [12]). Realizing this potential, PPF is incorporated into the IMBSS, which resulted in a conspicuous

improvement in the perceived speech quality with the annoying artifacts eliminated.

IMBSS Algorithm

The block diagram of the proposed IMBSS method is given in Fig.2, capturing the major stages involved. In the first stage, the noise signal is windowed and the magnitude spectrum is calculated using DFT. In the second stage the obtained magnitude spectrum is split into N linear non-overlapping frequency bands and pre-processing operation is performed individually in each band. In the third stage, over-subtraction and band subtraction factors are calculated for each band. The individual frequency bands are processed by subtracting corresponding noise spectrum from noisy spectrum, by an amount determined by the product of the above-obtained factors, to get the enhanced spectrum. In the next stage, this is passed on to a PPF (see Fig.4). The modified N frequency bands in every frame are recombined to get back the corresponding frames. Lastly, the time signal is obtained by using original noisy phase information and taking IDFT in the conventional way, i.e. OLA method.



The IMBSS method is identical with respect to step 1 of MBSS while certain modifications are carried out in the remaining steps, which are highlighted in the following. In order to calculate δ_i on a more methodical manner rather than empirical assignments, a mathematical equation connecting δ_i and a new parameter called Band Energy Content (BEC) is employed in IMBSS. A highlighting feature of the IMBSS method is that the frequency (band) dependent approach which was used for subtraction process alone (i.e. in MBSS), is carried forward to some other operations like preprocessing, noise flooring and noise updating.

Pre-Processing

While in the MBSS method (see Eqn. 10) averaging is done for the entire frame without any consideration for the frequency that a particular magnitude represented. But, here, this operation is carried out non-uniformly (weightage is given to frequency bands in the frame). Pre-processing is performed in each band as:

$$\overline{Y}_i(pL,k) = \sum_{l=-M}^{M} \sum_{i=1}^{N} W_{l,i} Y_i((p-l)L,k) \qquad (18)$$

where, p is the frame index, i is the frequency band number, N represents number of bands into which the whole spectrum is split. The averaging is done over M preceding and succeeding frames of speech centered on the (current) pth frame. $W_{l,i}$ is the band dependent weighting factor ($0 < W_{l,i} < 1$), used in a l^{th} preceding/succeeding frame and finally $\overline{Y}_i(pL,k)$ represents the ith frequency band of smoothed and averaged version of the noisy speech spectrum. It can be seen that the weighing factors are a function of frequency bands. The utility of this approach is that greater/lesser averaging can be performed at specific frequency bands as required in a particular situation. In our experiments, the isolated peaks in higher frequency, when converted back to time domain were more annoying than those in the lower frequencies. Thus, greater averaging was performed at higher frequencies.

Subtraction Method

Spectral subtraction is performed independently in each band as in Eqn.11. Calculation of α_i is identical to the way done in MBSS (Eqn.13).

Next, we explore an effective and adaptive method to obtain the values for δ_i. This is the central theme of the IMBSS method. The logic adopted here is to calculate δ_i in similar lines with that of α_i. α_i is dependent on the parameter SNR_i and thus it becomes imperative to find a similar parameter that rightly influences δ_i. After elaborate experimentation/observation, it is found that such a parameter can be defined. Now the obvious question that arises is that on what basis does one define such a parameter and the actual relationship between the parameter and δ_i, which could be intricately subtle. The rest of the section is devoted to explain these issues.

Since δ_i is a frequency-dependent factor, it would be appropriate that this parameter give a measure of the amount of relative speech energy present in each band of the noisy speech frame. Also, this parameter should respond to variations in noisy (speech) spectra so that the subtraction process is adaptive. We refer to this parameter as Band Energy Content (BEC) in this paper and is defined for a given noisy speech frame by:

$$\text{Band Energy Content} = \frac{\text{Energy in } i^{\text{th}} \text{ band}}{\text{Energy in the entire frame}} \tag{19}$$

That is,

$$BEC_i = \frac{\sum_{k=b_i}^{e_i} |\overline{Y}_i(k)|^2}{\sum_{i=1}^{NOB}\sum_{k=b_i}^{e_i} |\overline{Y}_i(k)|^2} \tag{20}$$

where, NOB represents the number of bands into which the spectrum is divided and b_i, e_i denote the beginning and ending frequencies of ith frequency band.

Band energy content gives a measure of how much speech and noise energy is present in a particular frequency band. A high value of BEC would mean a relatively more speech or noise is present in that band. From our exhaustive observations (for SNR > 0 dB), it is found that such a case is more likely due to speech than due to noise. In most of the cases, it was also found that a high value of BEC in a particular frequency band indicated that a relatively more speech energy was concentrated in that frequency band, with very less speech energy in the remaining frequency bands of the frame. Hence, as BEC increases it indicates more speech is present in that band and less amount of noise spectrum subtraction is needed, implying lower δ_i values and vice versa. In short, as the value of band energy content increases, the value of δ_i must decrease.

To formulate a relationship between the δ_i factor and BEC, the performance of the system was observed for 10 different speech sentences. For different values of BECi, δ_i was swept from 0-4 in steps of 0.1 units and the corresponding output (i.e. IS distance) was noted. The best sets of δ_i values, based on the IS distance, were plotted against the corresponding BEC value (BEC along x-axis). Using the best-fit piece-wise linear model, the graph of Fig.3 is obtained, with the corresponding equation given in Eqn.21. The graph exhibits a general non-increasing trend, except in the region 0.45 to 0.55, where a sharp discontinuity was observed.

Fig.3 Piece-wise linear variation of δ_i with BEC

$$\delta_i = \begin{cases} 2 & BEC \le 0.25 \\ -7.5\,(BEC) + 3.875 & 0.25 < BEC \le 0.45 \\ 0.5 & 0.45 < BEC \le 0.50 \\ 0.8 & 0.50 < BEC \le 0.55 \\ -7.5\,(BEC) + 4.925 & 0.55 < BEC \le 0.63 \\ 0.2 & 0.63 < BEC \le 1 \end{cases} \tag{21}$$

Spectral Flooring

Due to the error in computing the noise spectrum, we may have some negative values in the modified spectrum. The negative values obtained in the Eqn.11 were floored as:

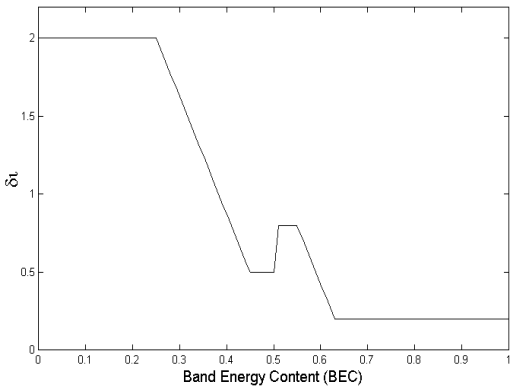
$$|\hat{S}(k)|^2 = \begin{cases} |\hat{S}(k)|^2 & \text{if } |\hat{S}(k)|^2 > 0 \\ \beta_i |\hat{D}(k)|^2 & \text{otherwise} \end{cases} \tag{22}$$

where, β_i is band-specific spectral floor parameter that can be set appropriately in each band, i.e. lower values in high frequency and vice versa.

Noise Spectrum Estimate/Update

The mean value of first five frames was used to obtain the initial estimated noise spectrum $\hat{D}(k)$. A statistical-model based voice activity detector (VAD), as in MBSS, is used to update estimated noise spectrum.

Accurate noise estimation is critical for the success of a spectral subtraction method. In order to come up with good noise estimator, the algorithm should be capable of reliably determining which portions of the signal are speech and which portions are not (therefore noise). For a low value of BEC in a particular frequency band of a noisy



speech frame, it implies that relatively a greater amount of noise energy is concentrated in that frequency band (from the previous discussion). This fact could be used to update the noise spectrum exactly in those frequency bands where the noise energies are relatively more. This operation is similar to the to the working of a VAD which detects those frames of noisy input speech where in the speech is absent (i.e. speaker is silent and mainly noise is present) and uses it to update the estimated noise spectrum. However, in VAD, the entire noisy speech frame is used for updating and in the suggested method, those particular frequency bands of a frame where noise energies are found high are used to update estimated noise. In the process, the method would increase the number of times the noise spectrum is updated and hence estimated noise spectrum would be more close to the actual noise spectrum. Whenever the BEC was more than 0.7 in a particular band of noisy speech frame, the noise spectrum was updated (this value was set based on the exhaustive observations carried out for different speech signals and noise levels). In short, compared to most VADs, the proposed method effectively updates the noise spectrum in a band-wise manner (instead of waiting for an entire frame where the speaker is silent), reducing the error between actual noise and estimated noise.

4. Perceptual Post Filtering (PPF)

Many SE algorithms suffer from the presence of musical noise, affecting their performance severely. In order to reduce the presence of musical noise a perceptual filter is used with the IMBSS algorithm (Fig 4).

It is well known that even though many conventional SE algorithms improve SNR of the noisy speech in the enhancement, this alone cannot guarantee significant increase in speech intelligibility, due to the quasi-stationary and other subtle properties of speech [11], [12]. To tackle this problem, researchers have been trying to incorporate the knowledge of human perceptual properties in the enhancement processing.

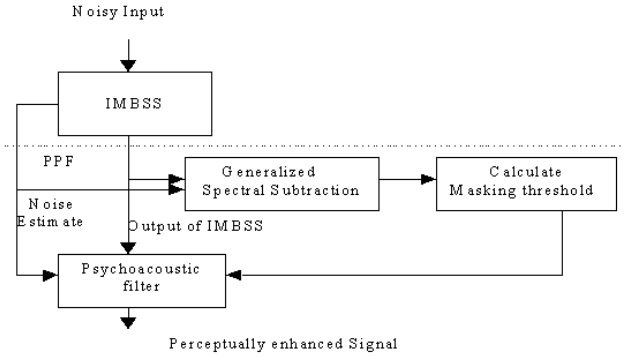


Fig 4 : Block diagram of IMBSS enhancement scheme with PPF

Virag [11] proposed a technique based on the masking properties of the human auditory system, i.e. the property that weak sounds are masked by simultaneously occurring stronger sounds. A masking threshold is calculated by modeling the frequency selectivity of the human ear and the masking property.

Considering the subtraction process described in Eq.6, the noise suppression can also be viewed as time-varying filtering process by rewriting the spectral subtraction method as:

$$\hat{S}(k) = H(k)\, Y(k) \tag{23}$$

where H(k) is a gain function represented by:

$$H(k) = \left[1 - \left(\left|\hat{D}(k)\right|^2 \Big/ \left|Y(k)\right|^2\right)\right]^{\frac{1}{2}}$$

$$H(k) = \left[\left(\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2\right) \Big/ \left(\left|Y(k)\right|^2\right)\right]^{\frac{1}{2}} \tag{24}$$

In this case, the modified spectrum is obtained by applying a time varying weight $H(k)$ to each frequency component. From Eqn.24, it can be deduced that the frequency dependent gain is a function of the noisy signal-to-noise ratio (NSNR) of each of the frequency components [12]. The enhanced time signal is synthesized as given in Eqn.8, using the original noisy phase portion.

Using the implementation of spectral subtraction the gain function can be calculated as [12]:

$$H(k) = \begin{cases} \left(1 - \alpha\left[\dfrac{\left|\hat{D}(k)\right|}{\left|Y(k)\right|}\right]^{\gamma_1}\right)^{\gamma_2} & \text{if} \quad \left[\dfrac{\left|\hat{D}(k)\right|}{\left|Y(k)\right|}\right]^{\gamma_1} < \dfrac{1}{\alpha+\beta} \\ \left(\beta\left[\dfrac{\hat{D}(k)}{Y(k)}\right]^{\gamma_1}\right)^{\gamma_2} & \text{else} \end{cases} \tag{25}$$

where, the over subtraction factor α and the spectral floor parameter β is a function of the masking threshold $T(k)$. The exponent γ_1 and γ_2 determine the sharpness of transition of $H(k)$. The masking threshold $T(k)$ is calculated by applying a spreading function across the critical bands of the speech spectrum [12]

The PPF was designed as per the steps mentioned in [6][12][13]. In [12], a perceptual filter was used with the signal-subspace speech enhancement algorithm. However, in this paper the PPF is modified to make it work in conjunction with the proposed algorithm, as shown in Fig.4.

The principle of PPF is as follows. A mathematical model of the human ear, based on auditory concepts, is formulated. The noisy spectrum is converted to bark scale (critical bands) as per excitation pattern of the human ear [12]. An estimate of the clean speech is obtained using SS. Estimated noise and clean speech spectrum is passed to the PPF. A masking threshold is calculated using the estimated noise and clean speech spectrum (see [12] for details). If the noise is below the masking threshold then the noise is not audible to the human ear and it is not processed. However, if the noise crosses the threshold, then it is audible and the noisy signal is multiplied by the gain function to bring the noise amplitude below the threshold. Thus, with the PPF only the noise audible to human ear is

suppressed. The PPF effectively removes the isolated peaks (musical noise) and smoothes the enhanced signal spectrum.

5. Results and Discussion

Exhaustive listening tests were carried out to evaluate the enhanced speech for different speech signals and noise scenarios. The latter refers to different noise levels as well as type: In the work, white noise, colored noise. For colored noise, speech-shaped noise as well as aircraft (cockpit) noise is used. Apart from listening tests, objective measures like SNR and IS distance are also used to measure the improvements achieved with the proposed method (lower the IS distance value the more close is the enhanced signal to the clean speech). In this paper, the results for white noise case alone are presented.

Improvements due to use of adaptive band-subtraction factors δ_i are tabulated in Table 1, for the output objective measures of mean SNR and IS distance, for five speech sentences. There is a marked increase in output SNR values and decline in IS distance.

Table 1: Effectiveness of using δ_i values(from Eq.18 instead Eq.10). Output objective measures are mean values of SNRs and IS dist. for 5 sentences.

Speech Condition	Mean Global SNR (dB)		Mean Segmental SNR (dB)		IS Distance	
Noisy Speech	5 dB	0dB	5 dB	0dB	5 dB	0dB
MBSS	8.51	5.50	4.44	2.74	1.19	1.40
MBSS using adaptive δ_i values	10.5	8.23	6.30	4.83	0.87	1.16

Spectrograms and time plots of noisy input with Additive White Gaussian Noise (AWGN) at 3dB, enhanced signal using MBSS, enhanced signal using IMBSS (without PPF) and enhanced signal using IMBSS with PPF are shown in Fig 5. It can be seen that some part of speech is distorted in MBSS output due to over-subtraction and inaccurate noise estimate. There is no speech distortion in IMBSS output, but small amount of musical noise is present. This is eventually removed by incorporating perceptual post filter. Table 2 compares the objective measures for MBSS and IMBSS (with and without PPF).

Apart from white noise, the proposed method is tested with colored noise conditions and the results showed improvements in both objective measures and in perceptual quality (through listening tests) over the earlier MBSS method. In fact, the performance is same or better than that of the subspace method with PPF [12].

CONCLUSIONS

This paper presented an improved version of Multiband Spectral Subtraction (MBSS) technique, referred to as IMBSS. The frequency-dependent (band-dependent) processing is considered for the other blocks of MBSS, apart from the subtraction process, like pre-processing, spectral flooring and noise updating. A parameter called Band Energy Content (BEC) for calculating band-subtraction factors is introduced in the algorithm. An effort has been made to update noise spectrum at certain frequency bands within a frame along with the conventional methods of noise updating during silent frames. The band subtraction factor is adapted using methodically quantified values based on BEC so as to counteract effectively any variations of speech/noise spectral properties. The Perceptual Post-Filter based on auditory masking is designed to operate along with the IMBSS algorithm. The PPF effectively removed almost entirely the presence of musical noise. The IMBSS algorithm with perceptual post filtering outperforms the MBSS algorithm (both in terms of objective measures and subjective listening tests), for white and colored noise situations. It provides a greater control of noise removal properties in different frequency bands. The overall system removes the presence of musical noise entirely, preserving the speech components with minimal distortions and restores the naturalness of speech.

REFERENCES

[1] J.S.Collura, "Speech enhancement and coding in harsh acoustic environment," in Proc. IEEE Workshop on Speech Coding, Porvoo, Finland, 1999, pp. 162-164.

[2] H. Levitt, "Noise reduction in hearing aids: An overview", Journal of Rehabilitation Research and Development, vol. 38, No. 1, January/February 2001.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust., Speech, Signal Process., vol.27, pp. 113-120, Apr. 1979

[4] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procs., pp. 208-211, Apr. 1979.

[5] S.Kamath, "A Multi-Band Spectral Subtraction Method for Speech Enhancement", Master Thesis, Univ. of Texas, 2001.

[6] Thomas F. Quatieri, "Discrete Time Speech Processing, Priniciples and practice", Pearson Education, 2002

[7] J. Deller Jr., J. Hansen and J. Proakis, "Discrete-Time Processing of Speech Signals",NY:
IEEE Press, 2000.

[8] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in

cars," Speech Communication, Vol. 11, Nos. 2-3, pp. 215-228, 1992.
[9] C. He and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation," ICASSP, vol.2, pp. 793-796, 1999.
[10] P.Sovka and P.Pollak, "The study of speech/pause detectors for speech enhancement methods," in Proc. 4th Eur. Conf. speech communication Technology EUROSPEECH'97, Rhodes, Greece, 1997.
[11] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, vol.7, pp 126-137, March 1999.
[12] M.Klein "Signal Subspace Speech Enhancement With Perceptual Post-Filtering," Master Thesis, McGill Univ. Montreal, Canada, May 2002.
[13] Z. Goh, K.Tan and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," IEEE Trans. Speech Audio Procs., vol. 6, pp. 287-292, May 1998.
[14] S. Quackenbush, T. Barnwell, and M. Clements, "Objective Measures for Speech Quality Testing," Prentice-Hall, 1988.
[15] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancements algorithms," Inter. Conf. on Spoken Language Processing, vol.7, pp. 2819-2822, Sydney, Australia, Dec.1998.
[16] Y. Ephraim, "Statistical-model-based speech enhancement systems," Proc. IEEE, vol.80, No.10, pp. 1526-1555, Oct.1992.

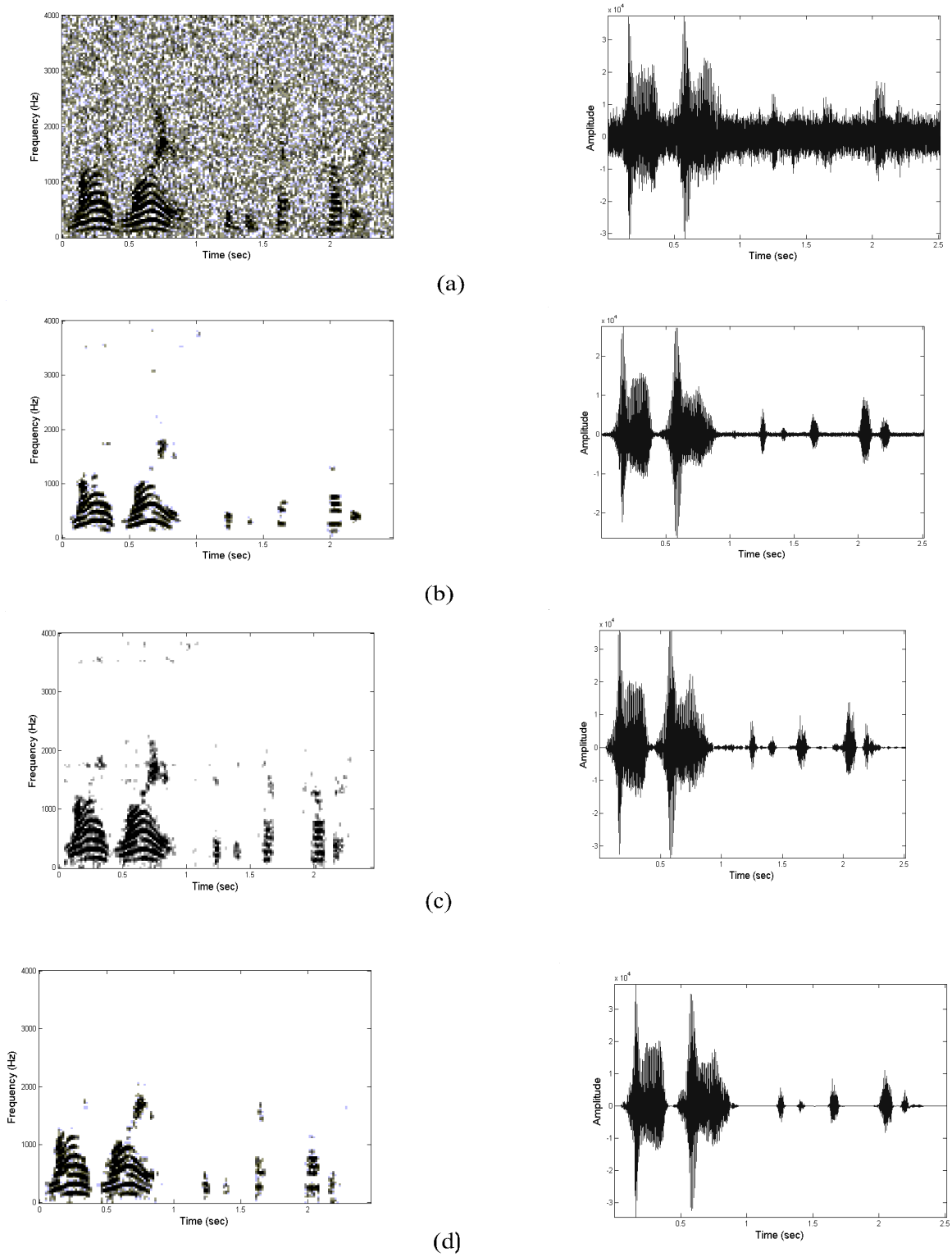


Fig -5: Spectrogram and time plot of (a) noisy (white noise), (b) enhanced signal using MBSS and (c) enhanced signal using IMBSS. (d) perceptually enhanced signal using IMBSS with PPF

Table 2: Comparison of global SNR, segmental SNR and mean IS distance for noisy speech, enhanced speech using MBSS, enhanced speech using IMBSS and enhanced speech using IMBSS with PPF

Speech condition	Global SNR	Segmental SNR	Mean IS Distance
Noisy Speech	3 dB	-1 dB	2.3
MBSS	7.23 dB	3.812 dB	1.19
IMBSS	8.68 dB	5.26 dB	0.90
IMBSS + PPF	8.67 dB	4.7 dB	0.98